

Introduction to Bioinformatics

**5. Multiple Sequence Alignment
and
Phylogenetic Trees**

Benjamin F. Matthews
United States Department of Agriculture
Soybean Genomics and Improvement
Laboratory
Beltsville, MD 20708
matthewb@ba.ars.usda.gov

What we will cover today

- Multiple Sequence Alignment (MAS)
- Motifs
- Phylogenetic Trees

What is Multiple Sequence Alignment (MSA)?

- An extension of a pairwise alignment
- Can be local or global
- The “inputs” are the same
 - A set of amino acid or nucleic sequences
 - Substitution (scoring) matrices
 - Gap penalties
- The objectives are similar: find an alignment of more than two sequences
- Discussed in earlier lecture

Multiple Sequence Alignment

Wheat	MSADKPSAYMLWLSNARESIKRENPDGIL
Rice	MKADKPSAYML - - - NARESI - - ENPDGRL
Soy	MPADKPSMFML - - - NPSESI - - NPDSARL

Why Do Multiple Sequence Alignment?

- Characterize protein families by identifying shared regions of homology
- Determine the consensus sequence of several aligned sequences
- Establish relationships and phylogenies
 - Clustering analysis
 - Structural modeling
 - Evolutionary analysis
- Use in a database search of protein families

Multiple Alignment programs

align several protein sequences

- ClustalW
 - <http://www.ch.embnet.org/software/ClustalW.html>
 - Multiple sequence alignment program
- T-Coffee
 - http://igs-server.cnrs-mrs.fr/~cnotred/Projects_home_page/t_coffee_home_page.html
 - Alignment program that often gives better results, especially when dealing with divergent sequences and long insertions

Multiple sequence alignment

- ClustalW
 - Does alignment and phylogenic tree
 - www.ebi.ac.uk/clustalw/index.html
- Dialign
 - Bibiserv.techfak.uni-bielefeld.de/dialign/
- Tcoffee
 - Igs-seerver.cnrs-mrs.fr/Tcoffee

MSA Algorithms

- As with the pairwise sequence comparisons, there are two types of multiple alignment algorithms
 - Optimal
 - Heuristic

Optimal MSA

- Extension of dynamic programming to multiple dimensions
- Exhaustive search
- Guaranteed to find an optimal score
- Need an n-dimensional matrix for scoring
- Computationally expensive
- Time complexity for pairwise comparisons is $O(m_1 * m_2)$; for multiple alignment should be $O(m^n)$
- Not feasible for $n > 10$ sequences of length $m > 200$ residues

Heuristic MSA

- Limit the exhaustive search
- Attempt to rapidly find a good, but not necessarily optimal alignment
- Most popular methods:
 - Tree alignments
 - Star alignments

Heuristic approaches to MSA

- Progressive global alignment starting from the most similar sequences: **CLUSTALW**
 - Pairwise alignment: calculate distance matrix
 - Neighbor joining: draw guide tree
 - Progressive alignment: align following guide tree
- Iterative methods: make initial crude alignment, then revise it: **DIALIGN**
- Alignment based on locally conserved patterns found in sequences in the same order: **BLOCKS, eMOTIF, GIBBS, MEME**
- Use statistical methods and probabilistic models of the sequences: **HMMER, SAM**

What is a motif?

- A short conserved region in DNA, RNA or protein sequence
- Corresponds to a structural or functional feature in proteins
- Shared by several sequences and can be generated by MSA
- Can be represented using position-specific scoring matrices

What is a profile?

- A position-specific scoring matrix, or matrix of scores representing a motif
- 22 columns, one for each of the 20 amino acids, and 2 for the penalties of opening and extending gaps
- The rows of the profile: aligned amino acid residues of a group of sequences
- Residues with the highest scores define a consensus

What is a protein family?

- A set of evolutionarily related proteins
- Members of a protein family may range from very similar to quite diverse
- Often share **domains**. Domain is a part of a protein (greater than a motif) that can fold and carry out a function independently.

Motif- and domain-oriented databases

- Secondary databases, small compared to GenBank
- Contain representations of conserved sequences shared by a sequence family
- Are primarily used for annotation of unknown sequences
- Examples: Pfam, Blocks, PRINTS, Prodom, PROSITE

Motifs and conserved domains

Pfam

- Protein FAMily (**Pfam**) is a large collection of multiple sequence alignments of sequence motifs or domains
- Made up of two parts: Pfam-A and Pfam-B
- **Pfam-A**: curated database of gapped profiles
- **Pfam-B**: generated automatically from sequences taken from the Prodom database that do not overlap with Pfam-A
- Use a Hidden Markov Model (HMM) to define domains or to align a set of sequences

Blocks

- Multiple sequence alignments without gaps that were used to construct the BLOSUM substitution matrices
- Generated automatically
- Correspond to the most conserved regions of a proteins
- Better used to identify protein sequence domains or families rather than identify motifs

PROSITE

- A database of sequence patterns (~motifs) associated with protein family membership
- Developed by largely manual process of seeking the patterns that best fit particular protein families
- Patterns may be useful in assigning distant homologs to sequence families
- PROSITE patterns are very short => may result in false positive occurrences in unrelated sequences

PRINTS

- A compendium of protein fingerprints
- A **fingerprint** is a group of conserved motifs used to characterize a protein family
- The motifs do not overlap (separated along a sequence)

Prodom

- **An automatically generated collection of protein domains**
- **Better described as a software tool to visualizes a protein's sequence domain structure**

Profile searches

- Numerical representations of multiple sequence alignments
- Depends upon patterns or motifs containing conserved residues
- Represent the common characteristics of a protein family
- Can find similarities among sequences with little or no sequence identity
- Allows for the analysis of distantly-related proteins

ProfileScan

- Search sequence against a collection of profiles and patterns
- Databases available
 - PROSITE profiles
 - PROSITE patterns
 - PfamA
 - PfamB
- <http://hits.isb-sib.ch/cgi-bin/PFSCAN>

Download

Profile Construction

APNIIIVATPG
 GCHIVIAATPG
 GVNICIAATPG
 GVNIIIGATPG
 KPNIIIVATPG
 KPNIIIAATPG
 KVQLIIAATPG
 KPNIVIAATPG
 APNIIIVATPG
 APNIIIVATPG
 GCHIVIAATPG
 GVNICIAATPG

- Which residues are seen at each position?
- What is the frequency of observed residues?
- Which positions are conserved?
- Where can gaps be introduced?

Position-Specific Scoring Table

Cons	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	X
G	17	18	0	19	14	-22	31	0	-9	12	-15	-8	15	30	9	6	18	14	1	-15	-22	11
P	-20	0	0	0	0	0	0	0	0	0	0	0	0	23	3	-2	12	11	17	-11	-8	1
N	5	24	-12	29	25	-20	8	32	-9	9	-30	-4	22	7	30	10	0	4	-8	-20	-7	27
E	-1	-12	6	-13	-13	33	-12	-13	63	-11	40	29	-15	-9	-14	-15	-6	7	50	-17	8	-11
V	3	-11	1	-11	-9	22	-3	-11	66	-9	37	30	-13	-3	-9	-13	-6	6	50	-19	2	-8
V	5	-9	9	-9	-9	19	-1	-13	57	-9	36	26	-13	-2	-11	-13	-4	9	58	-29	0	-9
A	54	18	12	30	17	-24	44	-4	-4	-1	-13	-8	12	19	9	-13	21	18	9	-19	-20	10
Q	40	20	20	30	20	-30	40	-10	20	30	-10	0	20	30	-10	-10	30	150	20	-60	-30	30
P	21	4	7	4	4	11	26	11	4	4	15	11	89	17	17	24	22	9	-80	-48	13	
G	-20	0	0	0	0	0	100	-30	-30	-10	-80	-30	60	30	20	-30	60	40	20	-100	-70	30

Download

Patterns

$[FY]-x-C-x(2)-[VA]-x-H(3)$

reads as:

Phe <i>or</i> Tyr	followed by
any amino acid	followed by
Cys	followed by
any two amino acids	followed by
Val <i>or</i> Ala	followed by
any amino acid	followed by
three His	

ProfileScan

- <http://hits.isb-sib.ch/cgi-bin/PFSCAN>
- Find all known motifs in a sequence

Protein sequence

```
LAQNPRSTLT PKARGFSRL L  
QIPEMASVSALAKYKLVFLG  
DQSVGKTSIITRFMYDKFDN  
TYQATIGIDFLSKTMYLED R  
TVRLQLWDTAGQERFRSLIP  
SYIRDSSVAVIVYDVASRQT  
FLNTAKWIEEV RTERGSDVI  
IVLVGNKTDLVEKRQVSIEE  
GEAKARELNVMFIETSAKAG  
FNIKALFRKIAAALPGMETL  
SSAKQEDMVDVNLKSTNGSA  
QSQPQSSGCAC *
```

Motif or conserved domain searches

- A pattern retained through evolution
 - not randomly changed by mutation
- Retained to help perform a specific function
- Domains can be found in databases
 - SMART
<http://smart.embl-heidelberg.de/>
 - Pfam
<http://www.sanger.ac.uk/Software/Pfam/>
 - COGs Clusters of Orthologous Groups
<http://www.ncbi.nlm.nih.gov/COG/>

Conserved Domain in Beta Globin

cd01040.1

globin

Inks:

Source: CDD
Taxonomy: cellular organisms
Pubmed: 5 links
Book: 3 book links
Proteins: cd01040 related
Related CDD: 4 links

Globins are heme proteins, which bind and transport oxygen. This family summarizes a diverse set of homologous protein domains, including: (1) tetrameric vertebrate hemoglobins, which are the major protein component of erythrocytes and transport oxygen in the bloodstream, (2) microorganismal flavohemoglobins, which are linked to C-terminal FAD-dependent reductase domains, (3) homodimeric bacterial hemoglobins, such as from *Vitreoscilla*, (4) plant leghemoglobins (symbiotic hemoglobins, involved in nitrogen metabolism in plant rhizomes), (5) plant non-symbiotic hexacoordinate globins and hexacoordinate globins from bacteria and animals, such as neuroglobin, (6) invertebrate hemoglobins, which may occur in tandem-repeat arrangements, and (7) monomeric myoglobins found in animal muscle tissue.

Feature 1: heme-binding site

Evidence: Structure: Ascaris hemoglobin with bound heme and oxygen molecule - [View structure](#) with Cn3D 4.1

Comment: Ascaris hemoglobin exhibits strong affinity to oxygen

Citation: PMID 7753786

Structure: Bovine deoxy-hemoglobin A with bound heme - [View structure](#) with Cn3D 4.1

Citation: PMID 9411160

Show Alignment

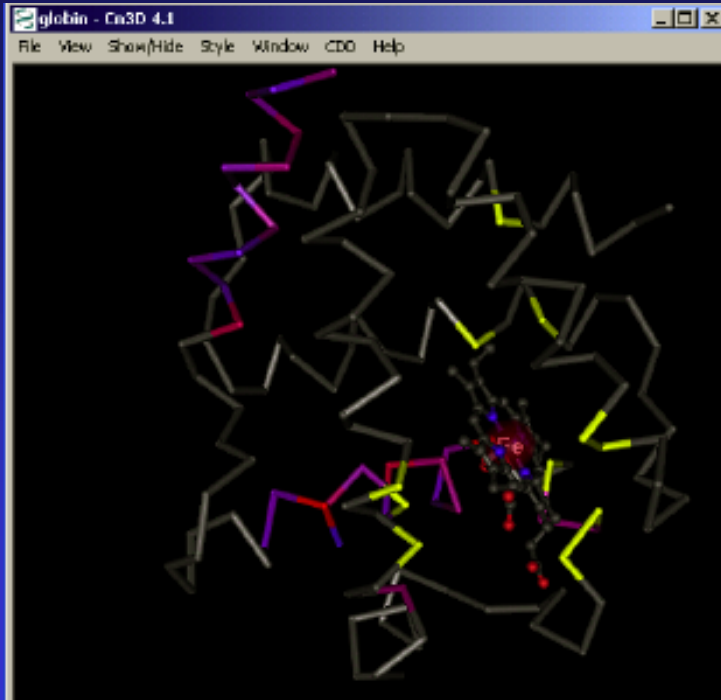
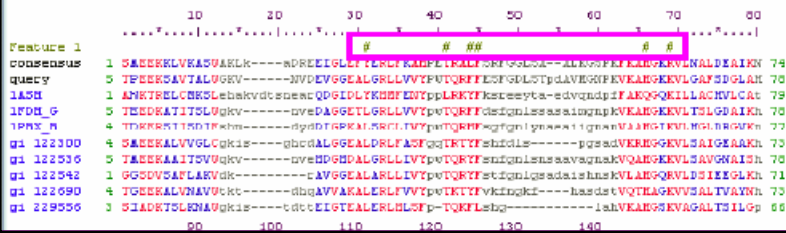
Format: Hypertext

Rows Display: up to 5

Color Bits: 2.0 bits

Type Selection: the most similar members


Feature Display: hemebinding site



File Edit View Favorites Tools Help


Back Forward Stop Reload Home Search Favorites Media Mail Print

Address http://hits.isb-sib.ch/cgi-bin/PFSCAN Go Links



Motif Scan in a Protein Sequence

(ProfileScan Server)



Motif scanning means finding all known motifs that occur in a sequence. This form lets you paste a protein sequence, select the collections of motifs to scan for, and launch the search. Some general [documentation](#) is available about the Prosite and Pfam collections of motifs. Another [document](#) deals with the interpretation of the match scores. You should consult the home pages of [Prosite](#) on ExPASy, [Pfam](#) and [InterPro](#) for additional information.

A pre-compiled list of matches is also available on our server ([Hits](#)). If your proteins of interest are already in the databases, the [Query by Protein](#) form is much faster, and the [Protein Hub](#) provides you a collection of tools that you might find useful.

Protein Sequence Input

Enter a protein sequence in RAW or FASTA format in this text area. Alternatively you can also paste a single sequence identifier, e.g.
sw:SLIT_DROME

```

L A Q N P R S T L T P K A R G
F S R L L

Q I P E M A S V S A L A K Y K
L V F L G

D Q S V G K T S I I T R F M Y
D K F D N

```

Reset
Clear

Motif Scan Parameters

☒ The Prosite profiles including the pre-released ones
☒ The Prosite patterns
☐ Prosite patterns that match most frequently


Database of motifs

Motif Scan in a Protein Sequence - Microsoft Internet Explorer

File Edit View Favorites Tools Help


Back Forward Stop Reload Home Search Favorites Media Mail Print

Address http://hits.isb-sib.ch/cgi-bin/PFSCAN_parser Go Links



Motif Scan in a Protein Sequence

result



Query

- Protein sequence:


```

>RAW_SEQUENCE
LAQNPRTLTPKAGFSRLQIPEMASVSALAKYKLVFLGDQSVGKTSIITRFMYDKFDN
TYQATIGIDFLSKTMYLEDRTVRLQLWDTAGQERFSLIPSYIRDSVAIVVDVASRQT
FLNTAKWIEEVRTERGSDVIIVLGNKTDLVKEKQVSIIEGEAKARELNVNFIETSAKAG
FNIALFRKIAAALPGNETLSSAKQEDMVDVNLKSTNGSAQSQPQSSGCACVYFIMCAPP
FCVIFVLCRFFSSSTSLYEKNKEYISDRGRGYKLCLENFCC

```
- Databases: Prosite patterns (weekly-updated), Prosite profiles (weekly-update), Pfam collection of hidden Markov models (weekly-updated)

...scanning for weekly_pattern
 ...scanning for weekly_profile
 ...scanning for weekly_pfam
 ...confirming pfam

Result

- Summary:

solhar Lecture 1 Adobe Acrobat 5.0.5 Motif Scan in a Protein Microsoft PowerPoint Adobe Acrobat

2 Motif Scan in a Protein Sequence - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://hits.isb-sib.ch/cgi-bin/PFSCAN_parser

Result

- Summary:
 - ?pfam.ATP_BIND_1 pos. 38 - 196 E-value=0.73
 - ?pfam.ARF pos. 21 - 194 E-value=3.3e-05
 - ?pfam.GTP_EFTU pos. 36 - 204 E-value=4.1
 - ! pfam.RAS pos. 35 - 196 E-value=3.3e-83
- Match Location:

query	LAQNPSTLTFRKGTRELLQIPZHAJVSALAKYKLVLGGQVUGKTSIITRINVKITDNTYQATIGIDFLSKTNVLEDR
pfam.ATP_BIND_1	<----->
pfam.ARF	<----->
pfam.GTP_EFTU	<----->
pfam.RAS	<----->

query	TURLQLVDTAGQERFRSLIPSYIRDSFVAVIVYDVASRQTPLNTAKWIEEVKTERGSDVILVLGKRTDLVEKRQUSIEE
pfam.ATP_BIND_1	<----->
pfam.ARF	<----->
pfam.GTP_EFTU	<----->
pfam.RAS	<----->

query	GEAKAPELVMMPIETSAKAGTNIKALFRKIAAALPGMETLSSAKQEDHMDVNLKSTNGSAQSQPSGACVTFINCAP
pfam.ATP_BIND_1	<----->
pfam.ARF	<----->
pfam.GTP_EFTU	<----->
pfam.RAS	<----->

query	FCVLPVLCRFSTSLVERKEVISEDGGGVGLCLNFCC
-------	--------------------------------------
- Prosite patterns (weekly-updated):
no match.
- Prosite profiles (weekly-update):

File Edit View Favorites Tools Help

Address http://hits.isb-sib.ch/cgi-bin/PFSCAN_parser

pos.: 36-204
raw-score = -90.8
N-score = 6.713
E-value = 4.1

Elongation factor Tu GTP binding domain
[Pfam-site](#)
[InterPro](#)

status: !
pos.: 35-196
raw-score = 289.9
N-score = 89.808
E-value = 3.3e-83

pfam: RAS
Ras family
[Pfam-site](#)
[InterPro](#)

Pfam: Search DNA vs. Pfam - Microsoft Internet Explorer

Address: <http://www.sanger.ac.uk/Software/Pfam/dnasearch.shtml>

Pfam Protein families database of alignments and HMMs
Wellcome Trust Sanger Institute

Pfam: Search DNA vs. Pfam
Home Search by Browse by ftp Pfam Help

This form allows you to compare your DNA sequence against the whole of Pfam using the [Wise2](#) software package

Cut and Paste your DNA sequence here, fasta format

```

ATATGCTTCGTTACATTACCTGAACAGTATTGTTGCT
TCCTTTTGTCTTTTAACTCCTGTGTATATGTTCTCCAT
TATTCAACTGTGAAGTGAATCATCATGCTTCTAT
TTTGGTGGTTTAGAACCTGTTGTTGACCTCTGTCAT
GTTTTCAGAAATTGTCATGTCAAAAGTTGATTCTTATG
TTTAAAA
  
```

* Searches now use the Sanger Blast queues. Presently the email option is not available *
It takes around 2 minutes for a 1,000 bp sequence, and around 2 hours for a 80kb sequence, depending on how many matches you get in them and the load on the sanger centre system

Or: Select the sequence file you wish to use
Browse...

DNA sequence is: Human Genomic DNA

Submit Query Reset Example

Output options for alignments

- ☒ Parameters
- ☒ Pretty alignment
- ☒ Predicted peptide
- ☐ Summary output
- ☐ GFF output
- ☐ Parseable alignment output

Output for gene predictions

- ☒ Gene structure as readable ASCII format
- ☐ EMBL feature table format
- ☐ EMBL feature table format suitable for [Artemis](#)
- ☐ AceDB subsequence object
- ☐ Translation as fasta format protein
- ☐ cDNA as fasta format DNA

If you think there is something wrong with this form or its results, please email Expasy@pfam.sanger.ac.uk

Pfam: p450 - Netscape

Address: <http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00067>

Pfam Protein families database of alignments and HMMs
Wellcome Trust Sanger Institute

Home Keyword Search Protein Search Browse Pfam DNA Search Taxonomy ftp Help p450 dnasearch

p450

Accession number: PF00067

Cytochrome P450 [Add description](#)

Cytochrome P450s are involved in the oxidative degradation of various compounds. Particularly well known for their role in the degradation of environmental toxins and mutagens. Structure is mostly alpha, and binds a heme cofactor.

INTERPRO description (entry IPR001126)

The cytochrome P450 enzymes constitute a superfamily of haem-thiolate proteins. P450 enzymes usually act as terminal oxidases in multicomponent electron transfer chains, called P450-containing monooxygenase systems and are involved in metabolism of a plethora of both exogenous and endogenous compounds. P450-containing monooxygenase systems primarily fall into two major classes: bacterial/mitochondrial (type I), and microsomal (type II). All P450 enzymes can be categorised into two main groups, the so-called B- and E-classes: P450 proteins of prokaryotic 3-component systems and fungal P450nor (CYP55) belong to the B-class; all other known P450 proteins from distinct systems are of the E-class [[MEDLINE:91135827](#)].

QuickGO

PROCESS : electron transport ([GO:0009118](#))

Figure 1: 2bmh Oxidoreductase(oxygenase)

Key:

Domain	Chain	Start Residue	End Residue
p450	A	5	448
p450	B	5	448

The SwissProt/PDB mapping was provided by MSD

For additional annotation, see the [PROSITE](#) document [P00009.1](#) [[ExPASy](#)] [SRS-UK](#) [[SRS-USA](#)]

→ Alignments
→ Domain Structure

Constructing Phylogenetic Trees

Why use phylogenetics

- Determine closest relative of an organism
- Discover the function of a gene
 - Identify orthologous, well-characterized gene in another species
- Retrace the origin of a gene
 - Mutations, deletions, gene duplications, gain- or loss-of-function, inactivation

Important

- Data quality
 - Highly accurate multiple sequence alignment that contains properly chosen sequences

Types of genes

- Orthologs
 - Separated only by speciation
 - A common ancestor gives birth to two subgroups that slowly drift away to become distinct species
- Paralogs
 - A gene is duplicated. The resultant two genes slowly diverge in sequence
- Xenologs
 - Result from a lateral transfer of a gene from one organism into the genome of another organism

How do you know two genes
coming from two different
species are orthologs or paralogs?

No simple solution

Strategy

- 1) Choose a sequence from genome A
- 2) GLAST search sequence A against every sequence in complete genome of B
- 3) BLAST search returns sequence B as a top hit
- 4) BLAST search B against every sequence in genome A
- Is sequence B the top hit???
- This does not prove – but provides support for

Tips

- Avoid sequence fragments
- Avoid xenologs
- Avoid recombinant sequences (especially seen in viruses)
- Avoid very large complex families containing repeats
- Keep your data set small
- Add an out group to root your tree
 - ie. a sequence that you know is a member, but has diverged long ago from the rest of the set

How to improve your multiple sequence alignment for phylogenetics

- Remove gaps
 - Gaps cause problems

Wheat	MSADKPSAYMLWLSN	ARES	IKREN	PD	SGIL
Rice	MKADKPSAYML	- - -	NARESI	- -	ENPDSGRL
Soy	MPADKPSMFML	- - -	NPSESI	- -	NPDSARL



How to improve your multiple sequence alignment for phylogenetics

- Remove extremities of your multiple sequence alignment
- N-terminus and C-terminus tend to be poorly conserved
- Remove gap-rich regions
- Keep most informative blocks
 - Typically 20 to 30 amino acids long
 - Contains a few conserved positions

Tree software

- ClustalW
 - easy
- Phylip
 - sophisticated

Example: tree alignment of four sequences



- Compare all six pairs of sequences
- Define and compute distances between the sequences
- Then use cluster analysis
- The number of pairs of N segments = $N(N-1)/2 = 4(3)/2 = 6$

Clustal W

- Format is important
- Can past or upload sequences
- Each sequence must have a unique name
- No empty lines
- No white spaces
- No control characters
- Limited to 500 sequences or 10MD, whichever is smaller)

>Arabidopsis

MVMAGASSLDEIRQAQRADGPAGILAIGTANPENHVLQAEYHDYYFRITNSEHKT
DLKEKFKRMCDSKSTIRKRHMHLEEFLENPHMCAYMAPSLDTRQDIVVVEVPKL
GKEAAVKAKEWGPQPSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMM
YQQGCFAGGTVLRIAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSL VGQAL
FSDGAAALIVGSDPDTSVGEKPIFEMVSAAQTLPDSDGAIDGHLREVGLTFHLLKD
VPGLISKNIKSLDEAFKPLGISDWSLFWIAHPGGPAILDQVEIKLGLKEEKMRT
RHVLSEYGNMSSACVLFILDEMRRKSAKDGVAATTGEGLEWGVSFSGPGLSVETV
VLHSVPL

>soyCHS5

MVSVEIRQAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNSEHMTLKEK
FKRMCDSMIKKRYMYLNEEILKENPSVCAYMAPSLDARQDMVMEVPKLGKEA
ATKAKEWGPQPSKITHLIFCTTSGVDMPGADYQLTKLLGLRPSVKRYMMYQQGC
FAGGTVLRLAKDLAENNGARVLVVCSEITAVTFRGPTDTHLDSL VGQALFGDGA
AAVIVGSDPLPVEKPLFQLVWTAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLSK
NIEKALVEAFQPLGISDYNSIFWIAHPGGPAILDQVEAKLGLKPEKMEATRHVLSEY
GNMSSACVLFILDQMRKKSIEGLGTTGEGLDWGVLFSGPGLTVETVLRSVTV

>SoyCH6

MVSVEIRQAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNSDHMNLKEKF
KRMCDKSMIKKRYMYLNEEILKENPSVCAYMEPSLDARQDMVVVEVPKLGKEA
TKAKEWGPQPSKITHLIFCTTSGVDMPGADYQLTKLLGLRPSVKRYMMYQQGCF
AGGTVLRLAKDLAENNTGARVLVVCSEITAVTFRGPSDTHLDSL VGQALFGDGA
AAVIVGSDPLPAEKPLFELVWTAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLSKNI
QKALVEAFQPLGIDDYNSIFWIAHPGGPAILDQVEAKLGLKPEKMEATRHVLSEY
NMSSACVLFILDQMRKKSIEGLGTTGEGLEWGVLFSGPGLTVETVLRSVTV

ClustalW Submission Form

Clustal W is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms. [New users, please read the FAQ.](#)

[Download Software](#)

YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT	CPU MODE
<input type="text"/>	Sequence	interactive	full	single

KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
def	def	percent	def	def

MATRIX	GAP OPEN	END GAPS	GAP EXTENSION	GAP DISTANCES
def	def	def	def	def

OUTPUT		PHYLOGENETIC TREE		
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS
aln w/numbers	aligned	none	off	off

Enter or Paste a set of Sequences in any supported format:

[Help](#)

Microsoft Internet Explorer

Address: <http://www.ebi.ac.uk/clustalw/index.html#>

Guide Tree
Colours

(WORD SIZE)	LENGTH			
def	def	percent	def	def
MATRIX	GAP OPEN	END GAPS	GAP EXTENSION	GAP DISTANCES
def	def	def	def	def

OUTPUT		PHYLOGENETIC TREE		
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS
aln w/numbers	aligned	none	off	off

Enter or Paste a set of Sequences in any supported format

Help

```
>SoyCH6
MVSVEEIRKAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNS
DHMNLKEKFRKMKDKSMIKRRYHYNLEEILKENPSVCAYMEPSLD
ARQDHVVVEVPKLGKEAATKAIKEWGPQSKITHLIFCTTSGVDMF
GADYQLTKLLGLRPSVKRYHMYQQGCFAGGTVLRLAKDLAENNTGA
RVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAADVIGSDPLP
AEKPLFELVWTAQTILPDSEGAIDGHLREVGLTFHLLKDVPGILSK
NIQKALVEAFQPLGIDDTNSIFWIAHPGGPALDQVEAKLGLKPEK
MEATRHLVSEYGNMSSACVLFILDQMRKKSIENTLGGTTGEGLEWGV
LFGFGPLTVETVVLRSVTY
```

Upload a file:

If you plan to use these services during a course please contact us using the email below.
Please read the FAQ before seeking help from our support staff.

Internet

Microsoft Internet Explorer

Address: <http://www.ebi.ac.uk/cgi-bin/clustalw/result?tool=clustalw&jobid=clustalw-20050202-18582521&poll=yes>

SEQUENCE ANALYSIS

Help
General Help
Formats
Gaps
Matrix
References
ClustalW Help
ClustalW FAQ
Jalview Help
Scores Table
Alignment
Guide Tree
Colours

ClustalW Results

Results of search	
Number of sequences	3
Alignment score	6306
Sequence format	Pearson
Sequence type	aa
ClustalW version	1.82
JalView	<input type="button" value="JalView"/>
Output file	clustalw-20050202-18582521_output
Alignment file	clustalw-20050202-18582521.aln
Guide tree file	clustalw-20050202-18582521.dnd
Your input file	clustalw-20050202-18582521.input

To save a result file right-click the file link in the above table and choose "Save Target As".
If you cannot see the JalView button, reload the page and check your browser settings to enable Java Applets.

Applet jalview.ButtonAlignApplet started

Internet

http://www.ebi.ac.uk/cgi-bin/jobresults/clustalw/clustalw-20050202-18582521.aln - Microsoft Internet Explorer

Back View Favorites Tools Help

Address http://www.ebi.ac.uk/cgi-bin/jobresults/clustalw/clustalw-20050202-18582521.aln

CLUSTAL W (1.82) multiple sequence alignment

```

oyCH6      -----MVSVEEIRKAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNSDHNNELKEK 55
oyCH55     -----MVSVEEIRKAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNSDHNNELKEK 55
rabidopsis MVMAGASSLDEIRQAQRADGPAGILAIGTANPENHVLQAEYHDYFFRITNSHKTDLKEK 60
           *::***:***:***:*****:* * * * * *****:* .:****

oyCH6      FKRMCDKSMIKKRYMYLNEEILKENPSCAYMEPSLDARQDMVVVEVPKLGKEAATKAIK 115
oyCH55     FKRMCDKSMIKKRYMYLNEEILKENPSCAYMEPSLDARQDMVVVEVPKLGKEAATKAIK 115
rabidopsis FKRMCDKSTIRKHHMLTEEFLENPHMCAYMEPSLDTRQDIVVVEVPKLGKEAATKAIK 120
           ***** *:***:*:*:*:*:*****:*****:***:***:*****:*****

oyCH6      EWGQPKSKITHLIFCTTSQVDMFGADYQLTKLLGLRPSVKRYMHYQQGCFAGGTVLR LAK 175
oyCH55     EWGQPKSKITHLIFCTTSQVDMFGADYQLTKLLGLRPSVKRYMHYQQGCFAGGTVLR LAK 175
rabidopsis EWGQPKSKITHVVFCTTSQVDMFGADYQLTKLLGLRPSVKRLMHYQQGCFAGGTVLR LAK 180
           ******:***:*:*:*:*:*****:*****:***:***:*****:*****

oyCH6      DLAENNTGARVLVVCSEITAVTFRGPSDTHLDSLQALFGDGA AAVIVGSDP--LPAEK 233
oyCH55     DLAENNTGARVLVVCSEITAVTFRGPSDTHLDSLQALFGDGA AAVIVGSDP--LPVEK 233
rabidopsis DLAENNTGARVLVVCSEITAVTFRGPSDTHLDSLQALFGDGA AAVIVGSDPDTSVGEK 240
           ***** *****:*****:*****:*****:*****:*****

oyCH6      PLFELVWTAQTILPDSEGAIDGHLREVGLTFHLLKDVFPGLISKNIQKALVEAFQPLGIDD 293
oyCH55     PLFELVWTAQTILPDSEGAIDGHLREVGLTFHLLKDVFPGLISKNIQKALVEAFQPLGIDD 293
rabidopsis PIFENVSAQTILPDSDGAIDGHLREVGLTFHLLKDVFPGLISKNIQKALVEAFQPLGIDD 300
           *:***:*:*****:*****:*****:*****:*****:*****:*****

oyCH6      YNSIFWIAHPGGPAILDQVEAKLGLKPEKMEATHRVLSEYGNMSSACVLFILDQMRKSI 353
oyCH55     YNSIFWIAHPGGPAILDQVEAKLGLKPEKMEATHRVLSEYGNMSSACVLFILDQMRKSI 353
rabidopsis WNSLFWIAHPGGPAILDQVEIKLGLKEEKMRATHRVLSEYGNMSSACVLFILDEMRRKSA 360
           :*:*****:*****:*****:*****:*****:*****:*****:*****

oyCH6      ENGLGTTGEGLEWGVLFQFGPGLTVETVVLRSVTV 388
oyCH55     ENGLGTTGEGLEWGVLFQFGPGLTVETVVLRSVTV 388
rabidopsis ENGLGTTGEGLEWGVLFQFGPGLTVETVVLRSVTV 388
           *****:*****:*****:*****:*****:*****:*****:*****

```

Done

ClustalW - Microsoft Internet Explorer

Edit View Favorites Tools Help

Back View Favorites Tools Help

Address http://www.ebi.ac.uk/cgi-bin/clustalw/result?tool=clustalw&jobid=clustalw-20050202-19573904&poll=yes

EMBL-EBI
European Bioinformatics Institute

Home About EBI Research Services Toolbox Databases Downloads Submissions

SEQUENCE ANALYSIS

ClustalW Results

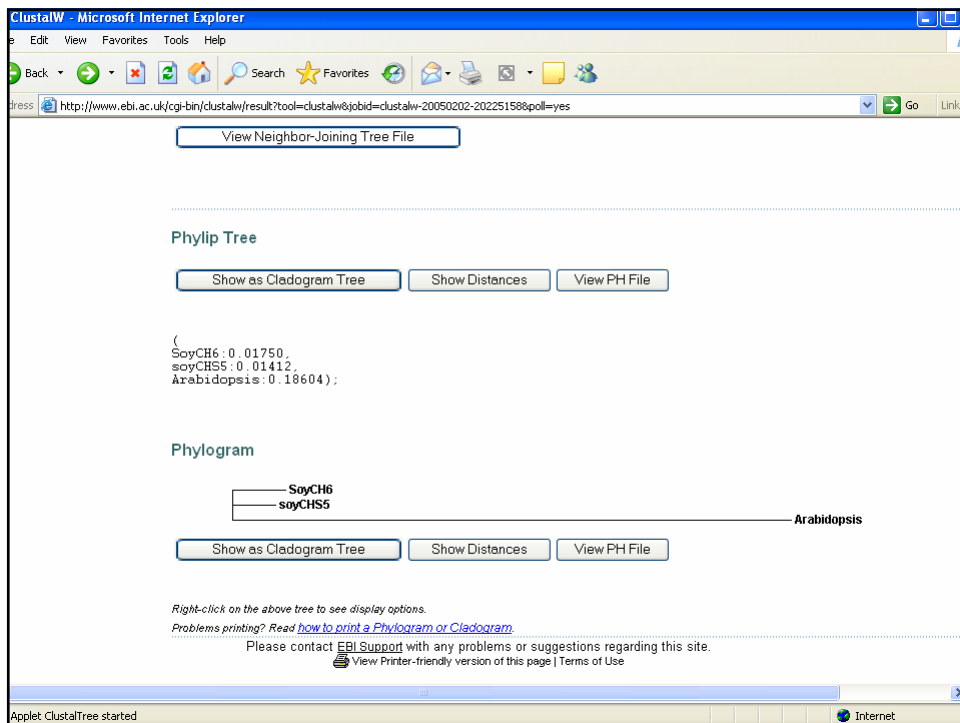
Results of search	
Number of sequences	3
Sequence format	Clustal
Sequence type	aa
ClustalW version	1.82
Output file	clustalw-20050202-19573904.output
Phylog tree file	clustalw-20050202-19573904.ph
Your input file	clustalw-20050202-19573904.input

SUBMIT ANOTHER JOB

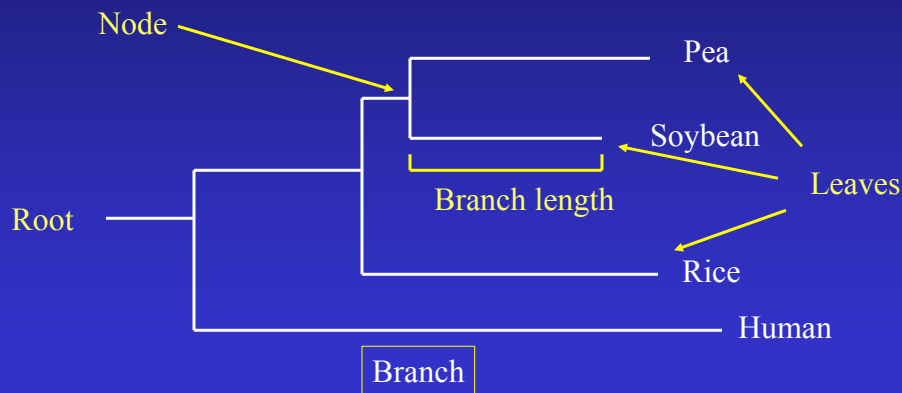
ClustalW Output

Get ClustalTree started

- Phylogram
 - Branch length represent real distances
- Cladogram
 - Branches indicate only branching order



Parts of a Phylogenetic Tree



Tree building

- Trees are also called **dendrograms**
- Nodes represent different organisms and links are used to show lines of descent
- Two basic types of questions about a tree:
 - a) its **topology**: how its interior nodes connect to one another and to the leaves
 - b) **distance between pairs of nodes**, which is an estimate of an evolutionary distance
- Tree may or may not have a root. A tree with root implies ancestral relationship between interior nodes

Phylogenetic tree evaluation

- How reliable phylogenetic tree is?
- One criterion: if different methods of tree construction give the same result, this is good evidence that the tree is reliable
- Another criterion (**bootstrapping**): data are randomly sampled from any position within MSA, and are built into new artificial alignments, which are then tested by tree building
- Third criterion (**jackknife**): similar to bootstrapping, but instead of generating new datasets with replacement, it re-samples the original data set by dropping one or more alignment positions in each replicate

What we learned

- Multiple Sequence Alignment (MSA)
- Motifs
- Phylogenetic Trees